Visualizing healthcare system dynamics in biomedical Big Data

HARVARD MEDICAL SCHOOL

PI: WEBER, GRIFFIN M

Grant Number: 1 U01 CA198934-01

Electronic health records (EHR) and administrative claims databases are transforming medical research by giving investigators access to data on millions of individual patients. Compared to manual paper chart review, these databases reduce the time and cost of clinical studies by orders of magnitude, enabling types of research that were unfeasible in the past. However, investigators often incorrectly treat EHR and claims data as simply big versions of clinical trials data. Yet, there are important differences: During clinicl trials, patient information is obtained and recorded in a standardized way and checked for accuracy and completeness. In contrast, EHR and claims are observational databases, which reflect not only the health of the patients, but also their interactions with the healthcare system For example, the date associated with a code for diabetes is when the physician made the diagnosis, not when the patient first developed the disease. These observations are influenced by the dynamics of the healthcare system-when physicians schedule visits with their patients, which tests physicians decide to order, what codes need to be recorded to get reimbursed for procedures, etc. By ignoring this dimension of the data or naively treating it as noise, investigators risk both misinterpreting the true patient pathophysiology and losing valuable information content. In prior work we showed that analysis of the ""healthcare system dynamics"" (HSD) dimension of observational databases can actually be more useful than the patient pathophysiology in predicting survival, selecting matched control cohorts, identifying healthy patients, and defining normal ranges of laboratory tests. Yet, conveying the concept of HSD to researchers and helping them use it effectively is difficult. Therefore, focusing on the topic area of Data Visualization, this proposal addresses this challenge of separating healthcare system dynamics from pathophysiology in observational databases, so that Big Data researchers can use both dimensions to generate new knowledge about patient health. To do this, we bring together informatics and data visualization experts who developed two widely adopted open source software platforms for querying clinical data repositories (Informatics for Integrating Biology and the Bedside, i2b2) and developing modular data analysis and visualization tools (Science of Science, Sci2). We will leverage these systems to perform three Specific Aims: (1) Create an extensible ontology for visualizing the HSD dimensions of biomedical Big Data. (2) Develop a prototype interactive visualization to enable investigators to study HSD in Big Data. The visualization will be simple and familiar to investigators, but innovative in that for the firs time HSD will be treated as its own informative component of the data. By literally placing HSD on its own dimension, the visualization will show investigators its value and teach them how to use it for research. (3) Demonstrate and evaluate the visualizations using three sources of biomedical Big Data: EHR data from two hospital systems in Boston with a total of 7 million patients and nationwide claims data from Aetna health insurance with 34 million patients. PUBLIC HEALTH RELEVANCE PUBLIC HEALTH RELEVANCE: Biomedical Big Data, such as electronic health records (EHR) and administrative claims are records of patients' interactions with the healthcare system; for example, the date of a diagnosis is when a physician entered the code into the EHR, not when the patient developed the disease. Most researchers are either unaware of the distinction or naively treat it as noise. However, the

proposed research will show, using a novel Data Visualization, that these subtle effects of the healthcare system on observational clinical data actually contain valuable information that could benefit biomedical research, clinical care, and health care policy.